KVM for IBM z Systems Performance

Martin Kammerer, Linux on z Systems Performance Evaluation

2016-07-20

Introduction

- *"Linux on z Systems is a true enterprise Linux solution!"*
  - *"efficiency at scale … fast data access … high utilization efficiency"*
- *"KVM for IBM z Systems provides open source virtualization for IBM z Systems and the LinuxONE platforms"*
  - *"advantage of the performance and scalability built into Linux"*
- High level performance tour through KVM host and virtual Linux servers
- Host / guest performance monitoring example

z Systems Technology for KVM Host (1)

- z Systems processor technology
  - Out-of-order processing
  - Transactional execution
  - Single instruction multiple data (SIMD)
  - Simultaneous multi threading (SMT not yet performance tested for KVM for IBM z Systems)
- z Systems network I/O (OSA Express)
- z Systems disk I/O
- z Systems crypto support
  - Clear key encryption with Central Processor Assist for Cryptographic Function (CPACF)

change in	affects	increases
Linux on z Systems	host and guests	overall throughput

z Systems Technology for KVM Host (2)

- Based on kernel level
  - RHEL 7 (kernel 3.10.0-...)
- Compiled for ...
  - Using z196 machine instruction set
    - march=z196
  - Instruction sequence optimized for zEC12
    - mtune=zEC12
- Future KVM releases could be created for only newer z Systems generations

change in	affects	increases
Linux on z Systems	host and guests	overall throughput

Processor Management

- Virtual servers are seen as processes in the KVM host process scheduler
  - Virtual CPUs appear as threads of the virtual server in the KVM host
  - Processor pinning in the guest is not recommended and at own risk
- Processor weight
  - A CPU weight can be configured for each virtual server (default 1024)
  - guest CPU share = (guest CPU weight)/(SUM of all guest's CPU weights)
  - CPU shares are overall values for each guest
    - Share per virtual CPU = guest CPU share / # virtual CPUs

change in	affects	increases
guest domain XML	all guests	processing power for this guest

Memory Management (1)

- Swapping
  - In general the alternative with poor performance
  - Swapping has not been heavily tuned in the past (workaround: increase memory)
  - Huge performance improvements implemented for z Systems
  - Swapping is the most inaccurate method to determine which guest pages are least important

change in	affects	increases
KVM host	all guests	memory space for all guests

Memory Management (2)

- Collaborative Memory Management Assist (CMMA)
  - Cooperative operation between KVM host and guest operating systems
  - z Systems hardware support allows sharing of guest's page state changes with the host
  - KVM host decides memory management based on page attributes provided by the guest operating system (takes only unused guest pages)
  - CMMA needs to be enabled for a guest
  - CMMA does not shrink the guest memory
  - Preventive action

change in	affects	increases
KVM host and guest OS	all guests	memory space for all guests

Memory Management (3)

- Ballooning with virtio_balloon driver
  - Cooperative operation between KVM host and guest operating systems
  - KVM host takes and returns pages from the guest to optimize memory usage
  - Balloon works like a memory consuming process in the guest
  - If balloon gets inflated it can lead to swapping in the guest
  - KVM host does not know the memory needs of the guests
  - Ballooning needs to be enabled for a guest
  - Do not use in combination with CMMA

change in	affects	increases
KVM host / guest domain XML	all guests	memory space for all guests

Memory Management (4)

- Kernel same-page-merging (KSM)
  - Shares memory pages between processes
  - KSM kernel thread scans memory periodically
  - Transparent for the guests
    - Enable by 'echo 1 > /sys/kernel/mm/ksm/run'
    - Watch by 'grep . /sys/kernel/mm/ksm/*'
    - Output example:

```
/sys/kernel/mm/ksm/full_scans:8
/sys/kernel/mm/ksm/pages_shared:68472
/sys/kernel/mm/ksm/pages_sharing:149009
/sys/kernel/mm/ksm/pages_to_scan:100
/sys/kernel/mm/ksm/pages_unshared:62702
/sys/kernel/mm/ksm/pages_volatile:1304
/sys/kernel/mm/ksm/run:1
/sys/kernel/mm/ksm/sleep_millisecs:20
```

change in	affects	increases
KVM host	all guests	memory space for all guests

Memory Management (4) cont.

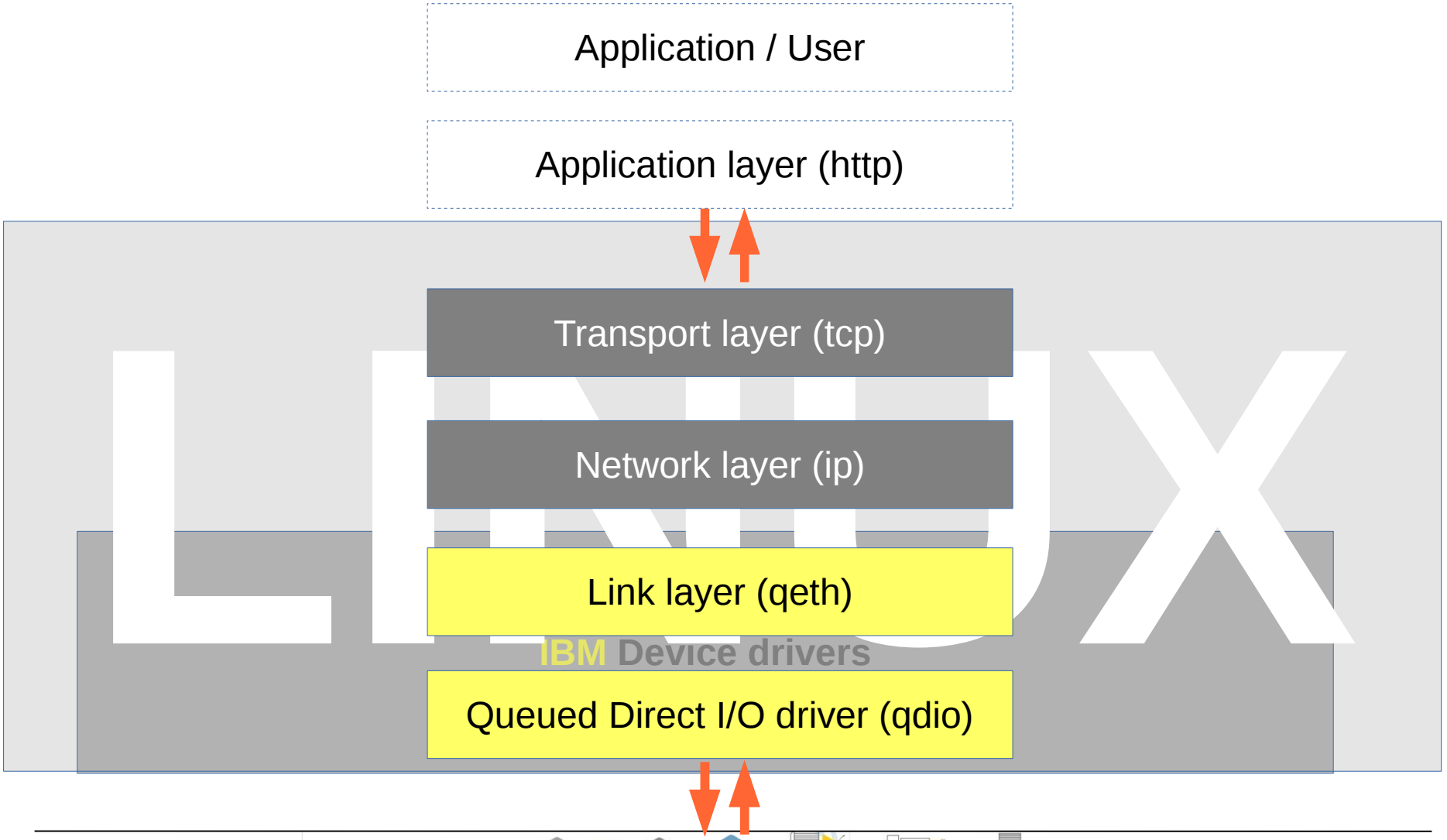
- Memory consumption in KVM host after booting 2 guests

```
[root@p10lp16]# free
              total        used        free      shared  buff/cache   available
Mem:        6165756     1234244     4718008       16820      213504     4818080
Swap:       7212140           0     7212140
```

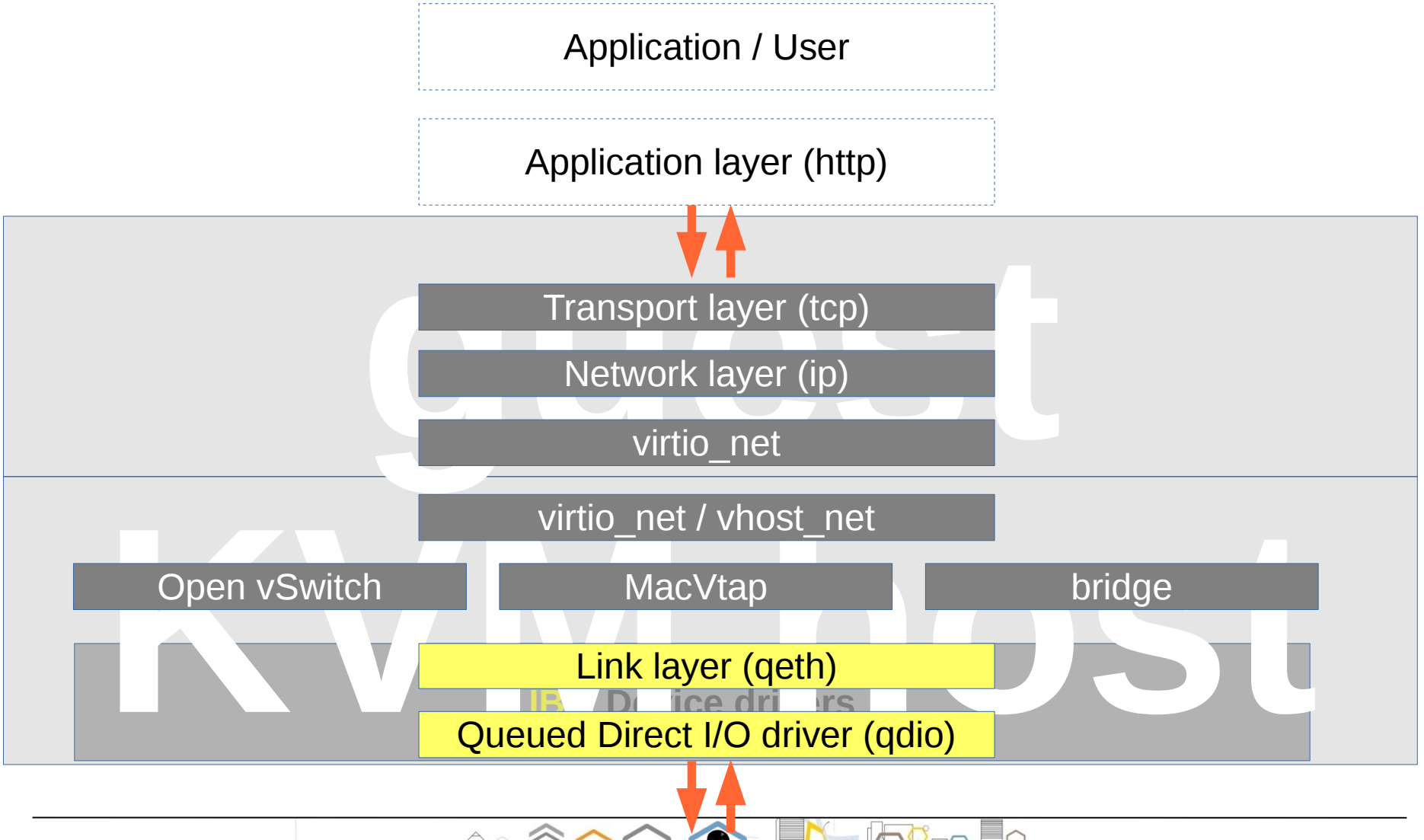
- Memory consumption in KVM host after activating ksm
  - Used memory reduced and free memory increased by 500 MiB

```
[root@p10lp16]# free
              total        used        free      shared  buff/cache   available
Mem:        6165756      734152     5146104       16820      285500     5285556
Swap:       7212140           0     7212140
```


Linux Network I/O, Flow Overview



Virtual Server Network I/O, Flow Overview



Linux Network Performance

- Network adapters
  - OSA Express 10 GbE, OSA Express GbE in layer 2 mode
- Parameter settings from Linux network tuning
  - Per interface
    - Appropriate maximum Transmission Unit (MTU) size
    - Device transmission queue length (txqueuelen)
    - Receive buffer count
    - Priority queuing (sharing mode)
    - Bonding
    - No TSO, GRO … with layer 2
  - System wide
    - Socket TIME_WAIT and reuse / recycle

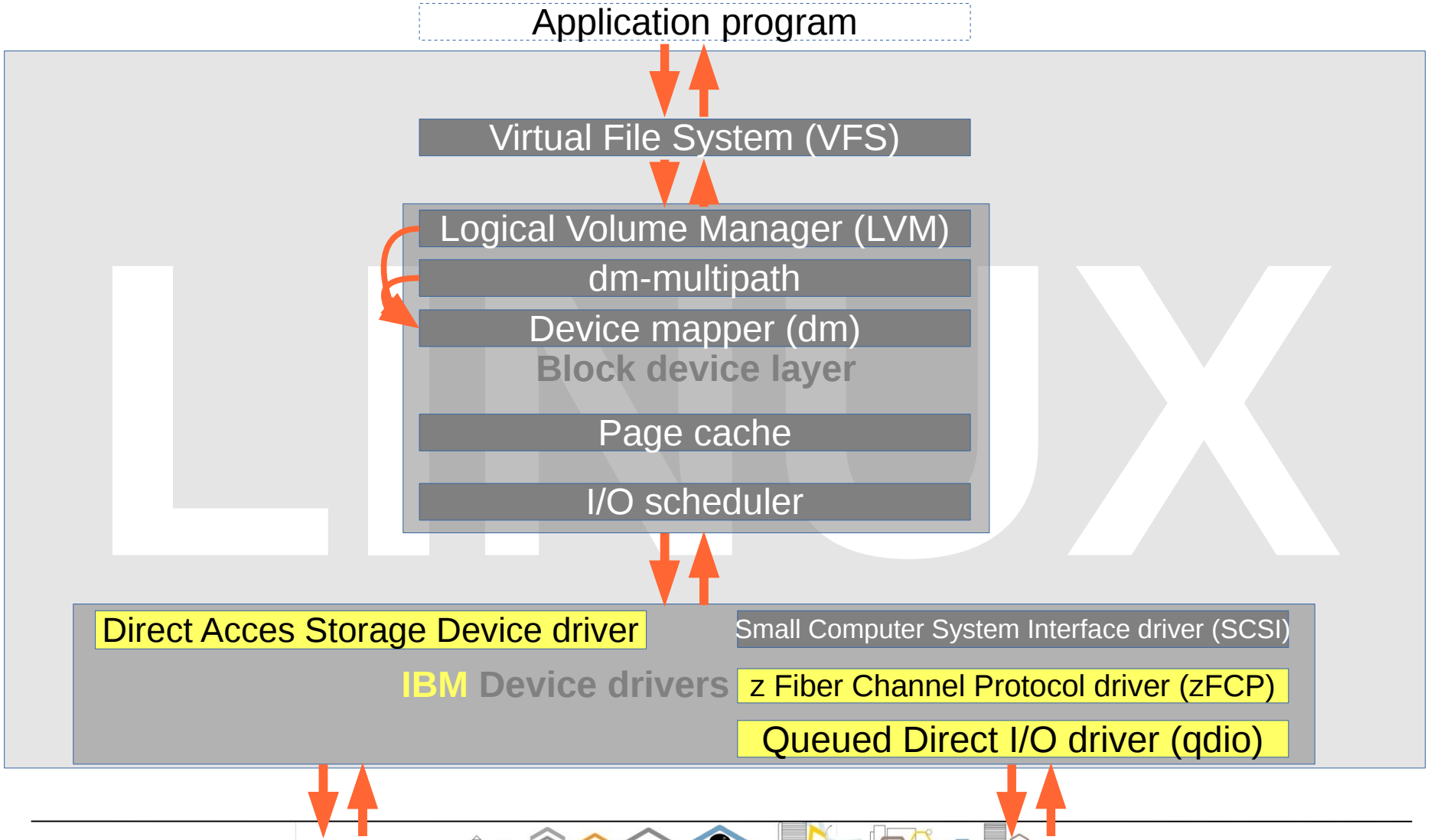
change in	affects	increases
KVM host	as described	overall throughput

KVM Networking Software Switches

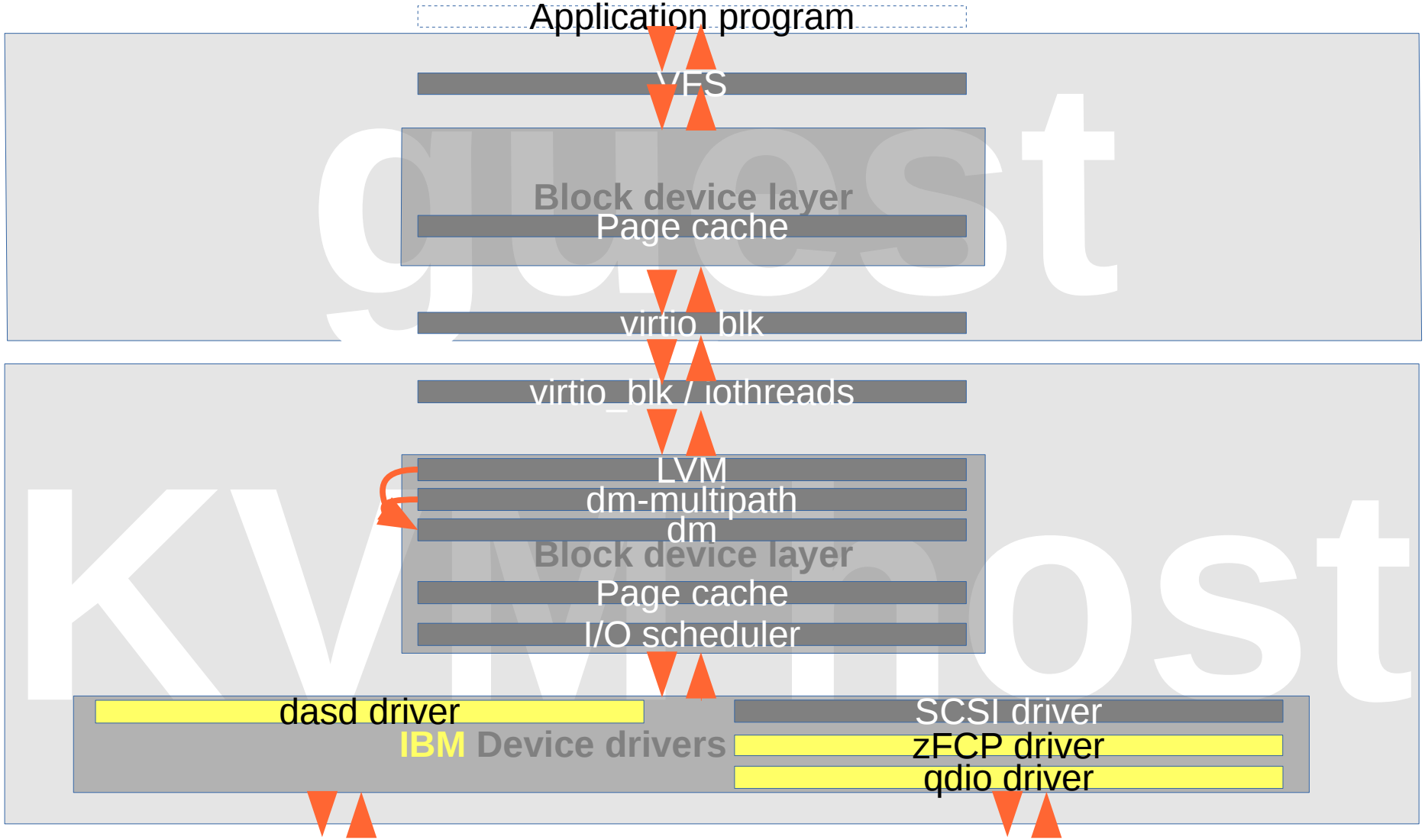
- Para-virtualized devices hide the real network devices for the guests
  - MacVTap
    - Best performance results so far
  - Open vSwitch
    - Has improved significantly over the past year or two
    - Working on some latency and processor consumption
    - Newer kernels have additional settings that improve transactional latencies
  - Linux bridge
    - Preliminary results show greater latency and processor consumption than Open vSwitch
    - Need more performance experience to make definitive statements

change in	affects	increases
guest domain XML	this guests	throughput of this guest

Linux Disk I/O, Flow Overview



Virtual Server Disk I/O, Flow Overview



KVM Host Disk I/O, Volume Options

- DASD driver for FICON/ECKD
  - Parallel Access Volumes (HyperPAV) boosts I/O
  - High Performance FICON (zHPF)
    - More throughput for random I/O (typically database)
  - Easy to configure, saves processor cycles
- zFCP and QDIO drivers for FCP/SCSI
  - Configure multipath devices of type multibus
    - Highest throughput
- Linear logical volumes allow an easy extension of the file system
- Striped logical volumes allow simultaneous I/O and load balancing

change in	affects	increases
KVM host	treated volumes	throughput of treated volumes

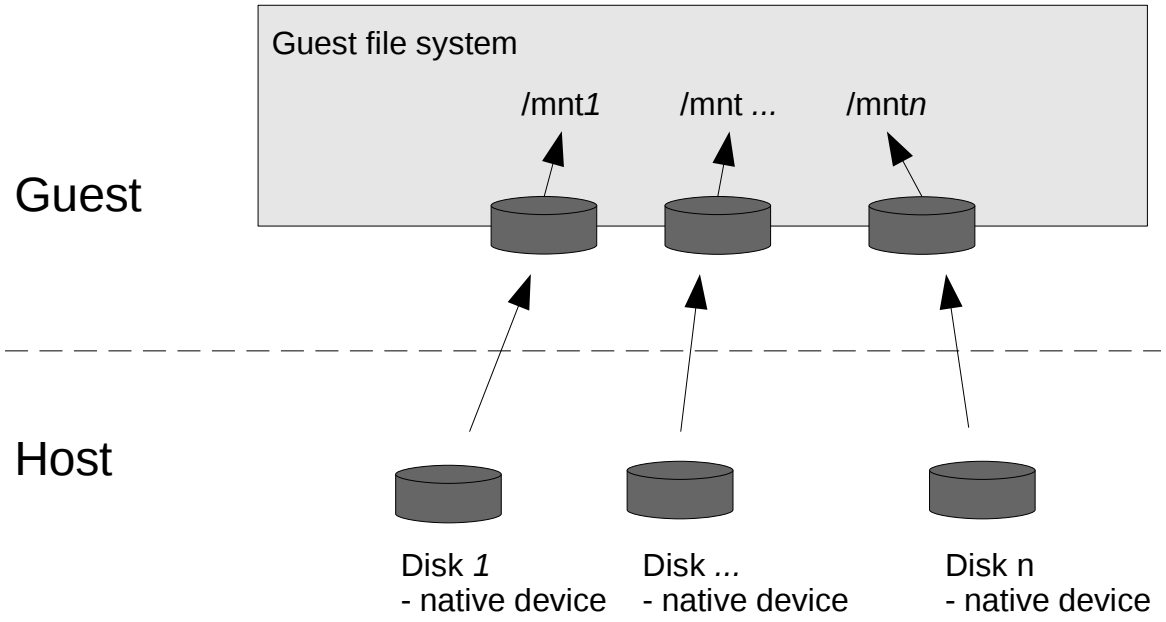
KVM Host Disk I/O, Storage Server

- For the host – storage server interconnect use many FICON Express channels
- Volume configuration in the storage server
  - Configure the volumes as storage pool striped volumes in extent pools larger than 1 rank, to use simultaneously
    - more physical disks
    - more cache and NVS
    - more device adapters
  - Provide alias devices (if possible HyperPAV) for FICON/ECKD volumes

change in	affects	increases
cabling, storage server	all hosts	throughput of all hosts

Disk Setup - Native Block Device

- The disk devices are 'owned' by the host as DASD or FCP device, but not mounted!
- The disk devices are propagated to the guest as independent virtio-blk data-plane devices.
- In the guest, each resulting virtio-blk device is partitioned and formatted with a file system and mounted.



Disk Setup - LV Based Block Device

- The disk devices are 'owned' by the host as DASD or FCP device
- Partitions are placed in a volume group and many logical volumes are created
- The logical volumes are propagated to the guest as independent virtio-blk data-plane devices
- In the guest, each resulting virtio-blk device is partitioned and formatted with a file system and mounted.

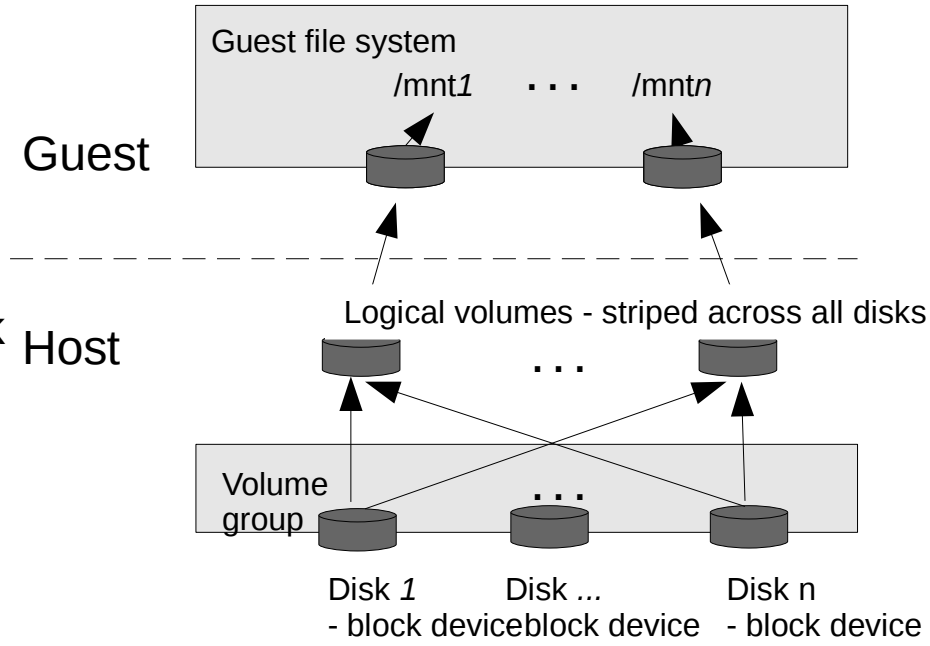


Image File Setup - One Image Per Disk

- The disk devices are 'owned' by the host as DASD or FCP device,
- Partitions are formatted and mounted on the Host
- The image file resides in the Host File system
- The image files are propagated to the guest as independent virtio-blk data-plane devices
- In the guest, each resulting virtio-blk device is partitioned and formatted with a file system and mounted

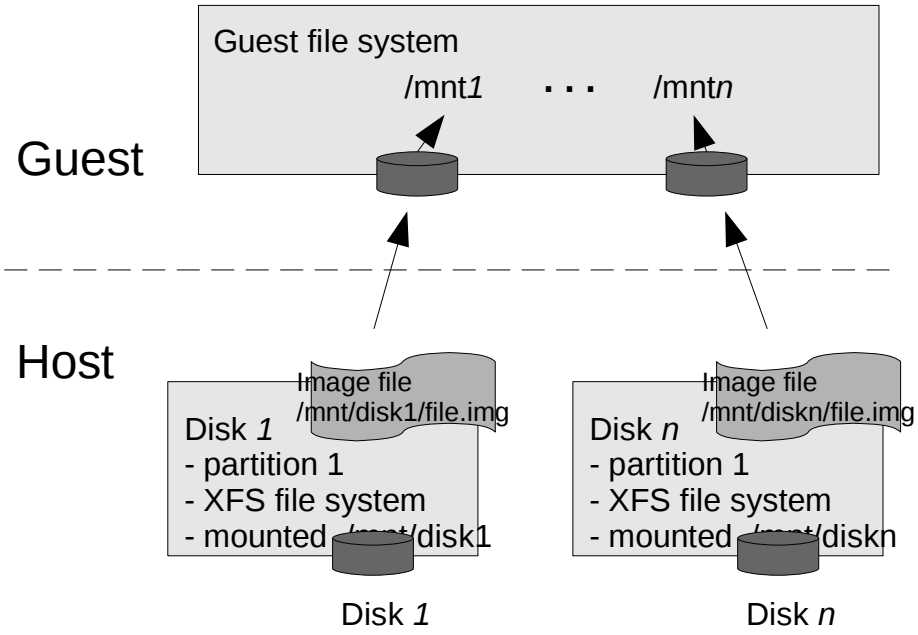


Image File Setup - LV Based Image Pool

- The disk devices are 'owned' by the host as DASD or FCP device,
- Partitions are placed in a volume group and one large logical volume is created
- The logical volume is formatted and mounted on the Host
- All image files reside in the Host File system in the logical volume
- The image files are propagated to the guest as independent virtio-blk data-plane devices
- In the guest, each resulting virtio-blk device is partitioned and formatted with a file system and mounted

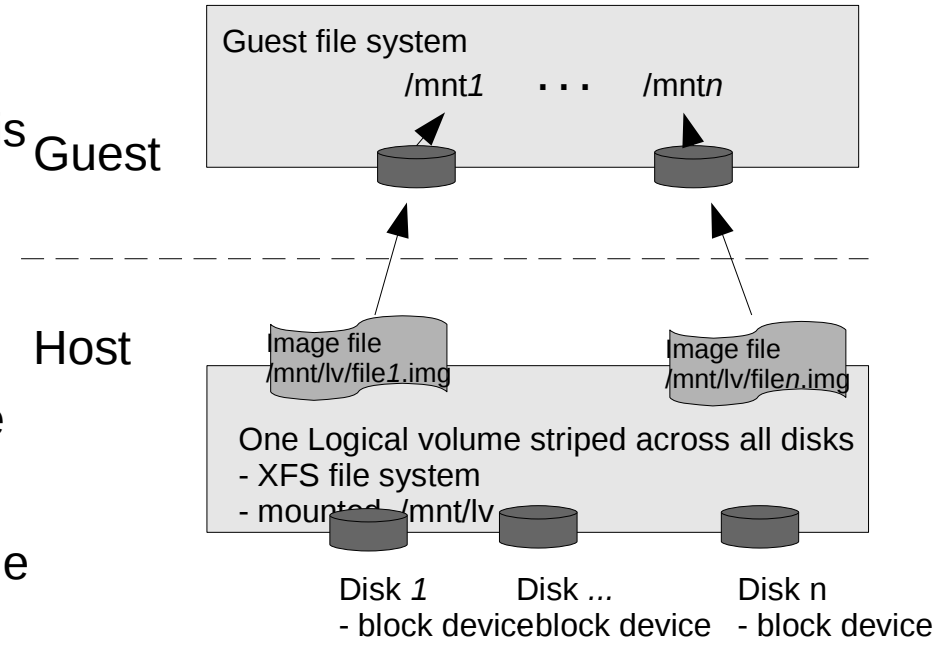
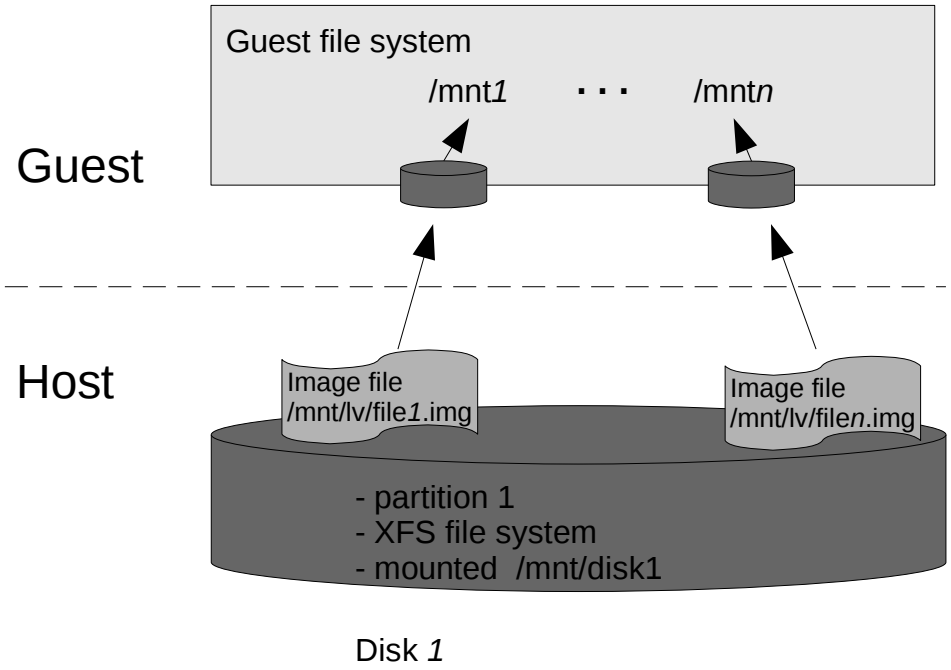


Image File Setup - Large Disk Based Image Pool

- The disk device is 'owned' by the host as DASD or FCP device,
- The disk is formatted and mounted on the Host
- All image files reside in the Host File system in the logical volume
- The image files are propagated to the guest as independent virtio-blk data-plane devices
- In the guest, each resulting virtio-blk device is partitioned and formatted with a file system and mounted.



KVM Host Disk I/O, Virtual Devices (1)

- Para-virtualized devices hide the real storage devices for the guests
- Devices presented to the guest could be
  - Block devices of DASD or SCSI volumes
    - For performance critical purposes
  - Disk image files
    - For performance uncritical purposes
    - Sparse files occupy only the written space (over commit disk space)
      - Tradeoff space consumption versus performance

change in	affects	increases
guest domain XML	this guests	throughput of this guest

KVM Host Disk I/O, Virtual Devices (2)

- KVM host and guest operating system maintain a page cache
  - Set cache=none in the device section of the guest configuration (recommended)
    - Disables the host's page cache for this device

change in	affects	increases
guest domain XML	this guest, KVM host	throughput of this guest

- Specify a quantity of iothreads in the guest configuration to enable simultaneous I/O
  - Performance uncritical I/O devices can share the same iothread
  - Performance critical I/O devices should have an individual iothread each
  - More than 32 iothreads will not improve throughput in case of many I/O devices

change in	affects	increases
guest domain XML	this guest, KVM host	throughput of this guest

Linux Guest Disk I/O, Performance Features

- Normal Linux disk I/O
  - Page cache helps to economize I/O accesses
  - Direct I/O bypasses the page cache
  - Async I/O prevents the application from being blocked until the I/O completes

change in	affects	increases
application in guest	this guest	throughput and CPU consumption of this guest

Monitoring Host and Guest - Preparation

- Using simple Linux tools
  - dstat
  - top
  - sadc / sar
- Forcing the tools to write their output to files
  - Allows comparison of collected data from the same time slot
  - Less resource consumption

Monitoring KVM Host (1)

```
[root@p10lp16 ~]# dstat -tclpymdn 1>hostdstat.out 2>&1 &
[root@p10lp16 ~]# cat hostdstat.out
----system---- ----total-cpu-usage---- ---load-avg--- ---procs--- ---system-- ------memory-usage----- -dsk/total- -net/total-
     time     |usr sys idl wai hiq siq| 1m   5m  15m |run blk new| int   csw | used  buff  cach  free| read  writ| recv  send
16-03 12:07:07|  1   0  99   0   0   0|0.27 0.32 0.26|  0   0 0.6| 295   579 |5270M 37.5M  136M  348G|  66k   16k|   0     0
16-03 12:07:08|  3   0  97   0   0   0|0.27 0.32 0.26|  0   0   0|1230  2421 |5270M 37.5M  136M  348G|   0     0 |   0     0

16-03 12:07:10|  3   0  97   0   0   0|0.25 0.32 0.26|  0   0   0|1227  2435 |5270M 37.5M  135M  348G|   0     0 |   0     0
16-03 12:07:11|  3   0  97   0   0   0|0.25 0.32 0.26|  0   0   0|1259  2495 |5270M 37.5M  135M  348G|   0    80k| 152B  276B
...
16-03 12:08:23| 23   0  77   0   0   0|0.27 0.30 0.26|  0   0 1.0|1383  2598 |5270M 37.6M  135M  348G|   0    16k|   0     0
16-03 12:08:24| 25   0  75   0   0   0|0.25 0.29 0.25|1.0   0   0|1382  2596 |5270M 37.6M  135M  348G|   0     0 | 332B  260B

16-03 12:08:26| 25   0  75   0   0   0|0.25 0.29 0.25|  0   0   0|1376  2591 |5270M 37.6M  135M  348G|   0     0 | 328B    0
16-03 12:08:27| 25   0  75   0   0   0|0.25 0.29 0.25|1.0   0   0|1343  2548 |5270M 37.6M  135M  348G|   0    32k|   0     0
...
16-03 12:09:30| 81   0  20   0   0   0|2.07 0.83 0.44|2.0   0   0| 955  1439 |5369M 37.7M  136M  348G|   0   104k|   0     0
16-03 12:09:31| 81   0  19   0   0   0|2.07 0.83 0.44|2.0   0   0|1000  1500 |5369M 37.7M  136M  348G|   0     0 |   0     0

16-03 12:09:33| 81   0  19   0   0   0|2.07 0.83 0.44|3.0   0   0| 943  1396 |5368M 37.7M  136M  348G|   0   216k|   0     0
16-03 12:09:34| 81   0  19   0   0   0|2.07 0.83 0.44|4.0   0   0| 865  1235 |5368M 37.7M  136M  348G|   0     0 |   0     0
```

Monitoring KVM Host (2)

```
[root@p10lp16 ~]# top -b -d 1 1>hosttop.out 2>&1 &
[root@p10lp16 ~]# cat hosttop.out
...
top - 12:09:32 up  3:17,  2 users,  load average: 2.07, 0.83, 0.44
Tasks: 106 total,   1 running, 105 sleeping,   0 stopped,   0 zombie
%Cpu(s):        ,  0.0 sy,  0.0 ni, 19.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem :                                            ,   243824 buff/cache
KiB Swap:  7212140 total,  7212140 free,        0 used. 36512832+avail Mem

   PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND

  6551 root      20   0  119584    900    744 S   1.0  0.0   0:00.28 sadc
     1 root      20   0    7164   4124   2460 S   0.0  0.0   0:00.59 systemd
     2 root      20   0       0      0      0 S   0.0  0.0   0:00.00 kthreadd
     3 root      20   0       0      0      0 S   0.0  0.0   0:00.01 ksoftirqd+
     4 root      20   0       0      0      0 S   0.0  0.0   0:00.00 kworker/0+
     5 root       0 -20       0      0      0 S   0.0  0.0   0:00.00 kworker/0+
     6 root      20   0       0      0      0 S   0.0  0.0   0:00.24 kworker/u+
     7 root      rt   0       0      0      0 S   0.0  0.0   0:00.00 migration+
     8 root      20   0       0      0      0 S   0.0  0.0   0:00.00 rcu_bh
     9 root      20   0       0      0      0 S   0.0  0.0   0:00.02 rcu_sched
    10 root      rt   0       0      0      0 S   0.0  0.0   0:00.00 migration+
    11 root      20   0       0      0      0 S   0.0  0.0   0:00.00 ksoftirqd+
    13 root       0 -20       0      0      0 S   0.0  0.0   0:00.00 kworker/1+
    14 root      rt   0       0      0      0 S   0.0  0.0   0:00.00 migration+
    15 root      20   0       0      0      0 S   0.0  0.0   0:00.01 ksoftirqd+
```

Monitoring KVM Host (3)

```
[root@p10lp16 ~]# /usr/lib64/sa/sadc -S ALL -F hostsadc.out 1 &
[root@p01lp16 ~]# sar -A -f hostsadc.out 1>hostsar.out 2>&1
[root@p10lp16 ~]# cat hostsar.out
Linux 3.12.49-11-default (p1016001)     16/03/16        _s390x_ (4 CPU)

12:07:25        CPU      %usr     %nice      %sys   %iowait    %steal      %irq     %soft    %guest    %gnice     %idle
...
12:09:32        all      0.00      0.00      0.25      0.00      0.00      0.00      0.00     80.50      0.00     19.25
12:09:32          0      0.00      0.00      0.00      0.00      0.00      0.00      0.00     80.00      0.00     20.00
12:09:32          1      1.00      0.00      0.00      0.00      0.00      0.00      0.00     80.00      0.00     19.00
12:09:32          2      1.00      0.00      0.00      0.00      0.00      0.00      0.00     80.00      0.00     19.00
12:09:32          3      0.99      0.00      0.99      0.00      0.00      0.00      0.00     79.21      0.00     18.81

12:07:25       proc/s   cswch/s
...
12:09:32         0.00   1422.00

12:07:25    kbmemfree kbmemused  %memused kbbuffers  kbcached  kbcommit   %commit  kbactive   kbinact   kbdirty
...
12:09:32    365112064   5675264      1.53     38640    139276   4789800      1.27   3913836    126372        68

12:07:25      runq-sz  plist-sz   ldavg-1   ldavg-5  ldavg-15   blocked
...
12:09:32            4       147      2.07      0.83      0.44         0
```

Monitoring KVM Guest (1)

```
[root@p1016001 ~]# top -b -d 1 1>guesttop.out 2>&1 &
[root@p1016001 ~]# cat guesttop.out
...
top - 12:09:32 up  3:16,  4 users,  load average: 3.95, 1.56, 0.84
Tasks: 129 total,   7 running, 122 sleeping,   0 stopped,   0 zombie
%Cpu(s):        ,  0.2 sy,  0.0 ni, 19.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.2 st
KiB Mem:                                              ,    17548 buffers
KiB Swap:        0 total,        0 used,        0 free.   192972 cached Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND

    1 root      20   0    6612   3976   2104 S 0.000 0.101   0:00.95 systemd
    2 root      20   0       0      0      0 S 0.000 0.000   0:00.00 kthreadd
    3 root      20   0       0      0      0 S 0.000 0.000   0:00.00 ksoftirqd+
    5 root       0 -20       0      0      0 S 0.000 0.000   0:00.00 kworker/0+
    7 root      rt   0       0      0      0 S 0.000 0.000   0:00.00 migration+
    8 root      20   0       0      0      0 S 0.000 0.000   0:00.00 rcu_bh
    9 root      20   0       0      0      0 S 0.000 0.000   0:00.15 rcu_sched
   10 root      rt   0       0      0      0 S 0.000 0.000   0:00.00 migration+
   11 root      20   0       0      0      0 S 0.000 0.000   0:00.00 ksoftirqd+
   13 root       0 -20       0      0      0 S 0.000 0.000   0:00.00 kworker/1+
```

Monitoring KVM Guest (2)

```
[root@p1016001 ~]# /usr/lib64/sa/sadc -S ALL -F guestsadc.out 1 &
[root@p1016001 ~]# sar -A -f guestsadc.out 1>guestsar.out 2>&1
[root@p1016001 ~]# cat guestsar.out
Linux 3.12.49-11-default (p1016001)     16/03/16        _s390x_ (4 CPU)

12:07:25        CPU      %usr     %nice      %sys   %iowait    %steal      %irq     %soft    %guest    %gnice     %idle
...
12:09:32        all     80.50      0.00      0.50      0.00      0.25      0.00      0.00      0.00      0.00     18.75
12:09:32          0     80.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00     20.00
12:09:32          1     80.00      0.00      1.00      0.00      0.00      0.00      0.00      0.00      0.00     19.00
12:09:32          2     79.21      0.00      0.00      0.00      0.99      0.00      0.00      0.00      0.00     19.80
12:09:32          3     82.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00      0.00     18.00

12:07:25       proc/s   cswch/s
...
12:09:32         0.00   2157.00

12:07:25    kbmemfree kbmemused  %memused kbbuffers  kbcached  kbcommit   %commit  kbactive   kbinact   kbdirty
...
12:09:32       338692   3580832     91.36     17548    173680   3535060     90.19   3375276    129416       848

112:07:25     runq-sz  plist-sz   ldavg-1   ldavg-5  ldavg-15   blocked
...
12:09:32            3       153      3.95      1.56      0.84         0
```

Monitoring KVM Host and Guest (1)

```
top - 12:09:32 up  3:17,  2 users,  load average: 2.07, 0.83, 0.44
Tasks: 106 total,   1 running, 105 sleeping,   0 stopped,   0 zombie
%Cpu(s):        ,  0.0 sy,  0.0 ni, 19.2 id,  0.0 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem :                                            ,   243824 buff/cache
KiB Swap:  7212140 total,  7212140 free,        0 used. 36512832+avail Mem

   PID USER      PR  NI    VIRT    RES    SHR S  %CPU %MEM     TIME+ COMMAND

  6551 root      20   0  119584    900    744 S   1.0  0.0   0:00.28 sadc
     1 root      20   0    7164   4124   2460 S   0.0  0.0   0:00.59 systemd

top - 12:09:32 up  3:16,  4 users,  load average: 3.95, 1.56, 0.84
Tasks: 129 total,   7 running, 122 sleeping,   0 stopped,   0 zombie
%Cpu(s):        ,  0.2 sy,  0.0 ni, 19.4 id,  0.0 wa,  0.0 hi,  0.0 si,  0.2 st
KiB Mem:                                              17548 buffers
KiB Swap:        0 total,        0 used,        0 free.   192972 cached Mem

  PID USER      PR  NI    VIRT    RES    SHR S  %CPU  %MEM     TIME+ COMMAND

    1 root      20   0    6612   3976   2104 S 0.000 0.101   0:00.95 systemd
```

Monitoring KVM Host and Guest (2)

```
12:07:25        CPU      %usr     %nice      %sys   %iowait    %steal      %irq     %soft    %guest    %gnice     %idle
...
12:09:32        all      0.00      0.00      0.25      0.00      0.00      0.00      0.00                 0.00     19.25
12:09:32          0      0.00      0.00      0.00      0.00      0.00      0.00      0.00                 0.00     20.00
12:09:32          1      1.00      0.00      0.00      0.00      0.00      0.00      0.00                 0.00     19.00
12:09:32          2      1.00      0.00      0.00      0.00      0.00      0.00      0.00                 0.00     19.00
12:09:32          3      0.99      0.00      0.99      0.00      0.00      0.00      0.00                 0.00     18.81
12:09:32        all                0.00      0.50      0.00      0.25      0.00      0.00      0.00       0.00     18.75
12:09:32          0                0.00      0.00      0.00      0.00      0.00      0.00      0.00       0.00     20.00
12:09:32          1                0.00      1.00      0.00      0.00      0.00      0.00      0.00       0.00     19.00
12:09:32          2                0.00      0.00      0.00      0.99      0.00      0.00      0.00       0.00     19.80
12:09:32          3                0.00      0.00      0.00      0.00      0.00      0.00      0.00       0.00     18.00

12:07:25       proc/s   cswch/s
...
12:09:32         0.00   1422.00
12:09:32         0.00   2157.00

12:07:25    kbmemfree kbmemused  %memused kbbuffers  kbcached  kbcommit   %commit  kbactive   kbinact   kbdirty
...
12:09:32                                     38640    139276   4789800      1.27   3913836    126372        68
12:09:32                                     17548    173680   3535060     90.19   3375276    129416       848

12:07:25      runq-sz  plist-sz   ldavg-1   ldavg-5  ldavg-15   blocked
...
12:09:32            4       147      2.07      0.83      0.44         0
12:09:32            3       153      3.95      1.56      0.84         0
```

Thank You!

Martin Kammerer
Manager Linux on z Systems Performance Evaluation

Research & Development
Schönaicher Strasse 220
71032 Böblingen, Germany

martin.kammerer@de.ibm.com

Linux on System z – Tuning hints and tips
http://www.ibm.com/developerworks/linux/linux390/perf/index.html

Live Virtual Classes for z/VM and Linux http://www.vm.ibm.com/education/lvc/

Mainframe Linux blog http://linuxmain.blogspot.com

Trademarks